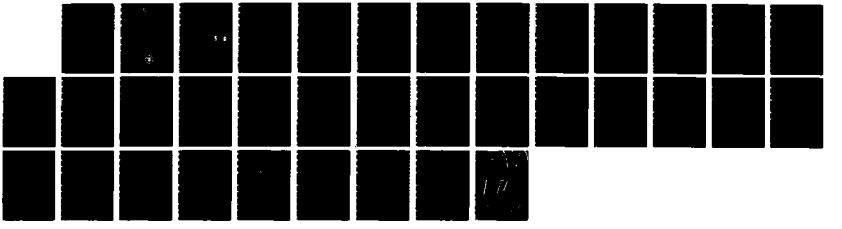
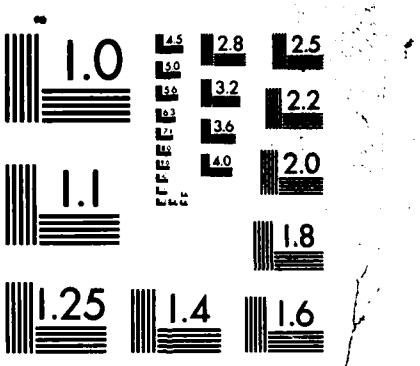
AD-A166 047 PARALLEL SOLUTION OF LINEAR SYSTEMS WITH STRIPED SPARSE MATRICES PART 2 S (U) PITTSBURGH UNIV PA INST FOR COMPUTATIONAL MATHEMATICS AND APP R MELHEM JAN 86 1/1

UNCLASSIFIED ICMA-86-92 N00014-85-K-0339 F/G 12/1 NL





14

INSTITUTE FOR COMPUTATIONAL MATHEMATICS AND APPLICATIONS

Technical Report ICMA-86-92 January, 1986

PARALLEL SOLUTION OF LINEAR SYSTEMS WITH STRIPED SPARSE MATRICES *

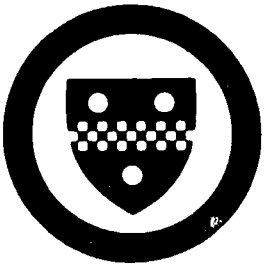
PART 2: Stiffness Matrices, A Case Study

by

Rami Melhem **

Department of Mathematics and Statistics

University of Pittsburgh



(14)

Technical Report ICMA-86-92 January, 1986

PARALLEL SOLUTION OF LINEAR SYSTEMS WITH STRIPED SPARSE MATRICES *

PART 2: Stiffness Matrices, A Case Study

by

Rami Melhem **

Institute for Computational Mathematics and Applications

Department of Mathematics and Statistics
University of Pittsburgh

* This work is, in part, supported under ONR contract N00014-85-K-0339

** On leave from the Department of Computer Science, Purdue University, West Lafayette, IN 47907.

PARALLEL SOLUTION OF LINEAR SYSTEMS WITH STRIPED SPARSE MATRICES *

Part 2: STIFFNESS MATRICES; A CASE STUDY

RAMI MELHEM **

Abstract

The stripe structures of stiffness matrices resulting from irregular domains covered by regular grids are analysed. It is proved that the non-zero elements in these matrices may be covered by very few stripes, and that these stripes may be non-overlapping, if the nodes of the grids are numbered appropriatly. The exact number of stripes, which is independent of the size of the problem, is derived for different types of grids, and different numbering schemes. The stripe structures of some irregular grids are also examined.

*) This work is, in part, supported under ONR contract N00014-85-K-0339.

**) On leave from the Dept. of Computer Science, Purdue Univ., West lafayette, IN 47907.

1. INTRODUCTION

Many techniques have been suggested for the efficient solution of sparse linear systems; they involve often highly irregular storage schemes and manipulation algorithms for the non-zero elements in the matrix [4]. Although, these techniques lead to very powerful sequential implementations (see e.g. [3]), they are not at all suitable for parallel architectures. In fact, parallel processing requires, in general, a rather regular pattern of computation in order to minimize data conflict and communication delays.

In Part 1 of this presentation (see [6]), a method is introduced for representing all non-zero elements of a sparse matrix in a stripe structure that provides, in some sense, a compromise between efficiency and regularity. More specifically, the stripe structure is shown to possess enough regularity to allow for the design of some efficient networks for the parallel manipulation of sparse matrices. Two networks, namely MAT/VEC for the multiplication of a matrix by a vector, and TRIANG, for the solution of triangular systems, are given as examples.

Very briefly, a stripe S_k of an nxn matrix A is a set of positions that contains at most one position, (i,j), of A for every row i; that is, $S_k = \{(i,\sigma_k(i)) : i \epsilon I \subseteq I_n\}$, where $I_n=\{1,\ldots,n\}$, and σ_k is a strictly increasing function. Two stripes S_k and S_q are ordered by $S_k < S_q$ if, for any i and j in the domains of σ_k and σ_q, respectively,

$$i \leq j \quad \text{implies} \quad \sigma_k(i) < \sigma_q(j) \tag{1}$$

A stripe structure of the matrix A is then defined as a disjoint union of stripes S_k, $k=1,\ldots,\pi$, which satisfies $S_k < S_{k+1}$, and contains all the non zero elements of A. More specifically, if $a_{i,j} \neq 0$, then, there should exist a unique k, such that $(i,j) \in S_k$. Also, the stripes $S_1,\ldots,S_\pi$, are said to be non-overlapping if

$$\sigma_k(i) \leq \sigma_{k+m}(i-m) \tag{2}$$

for any integers k, i and m such that $(i,\sigma_k(i)) \in S_k$ and $(i-m,\sigma_{k+m}(i-m)) \in S_{k+m}$. If the inequality in (2) is strict, then the stripes are called strictly non-overlapping.

The linear network MAT/VEC suggested in [6] for the multiplication of a matrix A by a vector x consists of π cells, where π is the number of stripes in the representation of A (called the stripe count). Every two consecutive cells k and k+1 in MAT/VEC are connected by two unidirectional communication links, where a link is regarded as a queue that may buffer data between cells k and k+1. One of the links is directed from k+1 to k and transmits the elements of the input vector x, and the other is directed from k+1 to k and transmits the elements of the result vector y=Ax. The network is data driven in the sence that the operation of each cell is initiated by the availability of its input.

In order to estimate the running time of MAT/VEC, it is

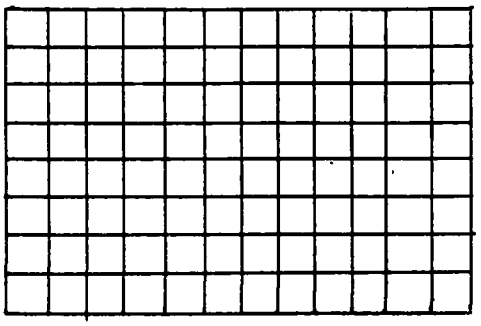
assumed that the execution alternates between two phases, namely a communication phase, and a processing phase. In [6], it is shown that if the input matrix to MAT/VEC has non-zero diagonal elements, and non-overlapping stripes, then no data conflict occurs, and the execution terminates in n global cycles, where a global cycle consists of a communication phase followed by a processing phase. The time for a processing phase is roughly that of one floating point operation, while the time for a communication phase depends on the slopes of the stripes of the matrix.

In this paper, we consider a major source of large sparse matrices; namely, finite elements and finite difference discretizations of partial differential equations (PDE). More specifically, we study the stripe structure of stiffness matrices that result from discretizations on irregular domains using regular grids. First, we specify in Section 2 the types of domains and grids used in the study. Then, in Section 3, we show that for matrices resulting from these types of grids, a stripe structure with very few stripes may be introduced, but the resulting stripes do, in general, overlap. In order to obtain non-overlapping stripes, we suggest , in Section 4, a multicolor numbering scheme that spreads the stripes within a matrix, and thus disengages any overlap between stripes. The multicolor numbering is shown, in Section 5, to decrease the maximum separation between stripes, which minimizes the number of data items that should be buffered, at any give time, on the

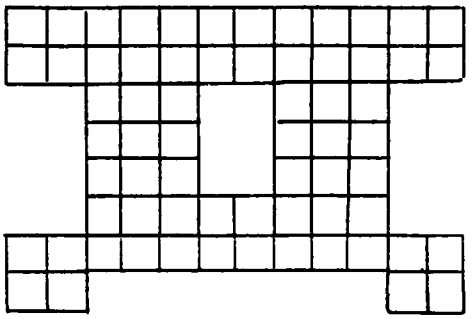
communication links of MAT/VEC. Finally, in Section 6, we estimate the execution time of MAT/VEC for some specific stiffness matrices.

2. PIERCED RECTANGULAR DOMAINS

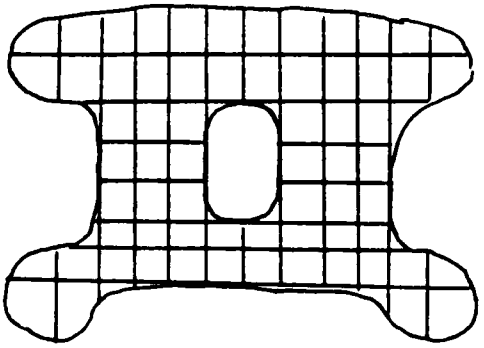
Let Q be a rectangular domain that is covered by a grid M_Q with lines parallel to the sides of Q. If we remove from Q any number of rectangular subdomains whose boundaries coincide with some lines in M_Q, then we obtain a new domain $\Omega \subset Q$, which we will call a pierced rectangular domain. The part of M_Q that covers Ω is denoted by M_Ω and is called a pierced rectangular grid (see Fig 1(b)).



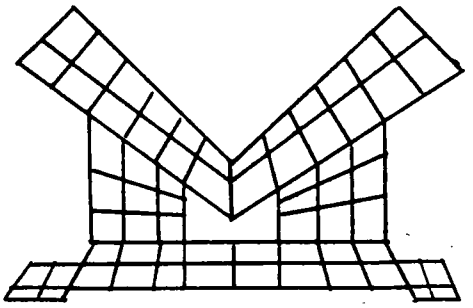
(a) a rectangular grid M_Q



(b) a pierced rectangular grid M_Ω



(c) an irregular domain covered by M_Ω



(d) an irregular grid M_D isomorphic to M_Ω

Fig 1 - Examples of finite elements grids

If D is an irregular domain, we may approximate D by a pierced rectangular domain and cover it by a pierced rectangular grid (see Fig 1(c)). Another alternative (usually

used in automatic grid generation), is to map D, isoparametrically [9], into a pierced rectangular domain Ω, cover Ω by a pierced rectangular grid M_Ω, and then map M_Ω back to a grid M_D that covers D (see Fig 1(d)). In this case, the zero pattern of the stiffness matrix that results from the discretization of a PDE on M_D is the same as that resulting from the discretization of the PDE on M_Ω. For this reason, we consider here only discretizations on pierced rectangular grids.

Given a pierced rectangular domain Ω covered by a gr. M_Ω that contains n_Ω nodes, let M_Q be a rectangular grid th: includes M_Ω and contains n_Q nodes, $n_Q > n_\Omega$. Each node in M may be identified by a unique number λ, $1 \le \lambda \le n_Q$, assigned to it by some numbering scheme (greek letters will be used to identify nodes in M_Q). On the other hand, the nodes in M_Ω may be renumbered such that each node $\lambda \epsilon M_\Omega$ is assigned a unique number $\ell = \nu(\lambda)$, $1 \le \ell \le n_\Omega$.

Definition 1: A renumbering of M_Ω is said to be deduced from the numbering of M_Q if the number $\ell = \nu(\lambda)$, assigned to any node $\lambda \epsilon M_\Omega$ is derived as follows:

$\ell = 0$

For $\lambda = 1,\ldots,n_Q$ Do

If $\lambda \epsilon M_\Omega$ Then { $\ell = \ell + 1$; $\nu(\lambda) = \ell$ }

Else { $\nu(\lambda)$ is undefined }

Clearly, the renumbering function ν satisfies the following relations:

$$\lambda > \mu \quad \langle\equiv\equiv\rangle \quad \nu(\lambda) > \nu(\mu) \tag{3.a}$$
$$\lambda > \mu \quad \equiv\equiv\rangle \quad \lambda-\mu \geq \nu(\lambda)-\nu(\mu) \tag{3.b}$$

The inverse ν^{-1} of the function ν will be used to map the number ℓ of any node in M_Ω into its identity $\lambda=\nu^{-1}(\ell)$ in M_Q. It is also useful to define a function which determines, for each node $\delta \notin M_\Omega$, the smallest node larger than δ that is in M_Ω. For uniformity, we define such a function for any $\delta \in M_Q$ as follows:

Definition 2: The function Next(δ) : $M_Q \rightarrow M_Q$ is defined by:

$$\text{Next}(\delta) = \begin{cases} \min\{\mu : \mu \geq \delta \text{ and } \mu \in M_\Omega\} & \text{if such } \mu \text{ exists} \\ n_Q & \text{otherwise} \end{cases}$$

Note that the minimum does not exist if every node $\mu \geq \delta$ is not in M_Ω, which may happen only if $\delta > \nu^{-1}(n_\Omega)$.[]

Without entering into the details of the generation of stiffness matrices, we just mention that the matrix A generated from the discretization of a PDE on M_Ω is an $n_\Omega \times n_\Omega$ matrix in which each row ℓ corresponds to a node $\lambda=\nu^{-1}(\ell)$ in M_Ω. The only non zero elements in row ℓ of A are those at positions (ℓ,m), where $\mu=\nu^{-1}(m)$ is a node that is a neighbor to node λ in M_Ω. The definition of neighboring nodes depends on the specific discretization used. For example, in finite element discretizations, two nodes are neighbors if there exists an element that contains the two nodes.

From the above discussion, it is clear that the scheme used to number the nodes determines the zero structure of

the matrix A. In the following subsections we consider two different numbering schemes. For ease of reference, we refer to the 5-points star finite difference discretization by FD_5, and to finite elements discretizations with 3-nodes triangles, 6-nodes triangles, 4-nodes rectangles, and 9-nodes rectangles by FE_3, FE_6, FE_4, and FE_9, respectively.

3. REGULAR NODE NUMBERING

A regular node numbering is one in which the nodes are numbered sequentially, column-wise or row-wise. We will consider only column-wise numbering and note that our results apply to row-wise numberings, as well.

Let M_Q contains H horizontal lines and W vertical lines, and identify each node in M_Q by the number assigned to it by the column-wise numbering of M_Q, that is, identify the node located at the intersection of the i^{th} horizontal line and the j^{th} vertical line of M_Q by the integer $(j-1)H+i$. It is easy to see that the column-wise numbering of M_Ω is the one deduced from the above numbering of M_Q. Let ν be the renumbering function introduced in Definition 1.

Depending on the specific discretization, we may introduce few functions that define the neighbors of each node λ in M_Q. For example, for FE_4 discretization, the following nine neighboring functions may be defined for each $\lambda \epsilon M_Q$ (see Fig 2):

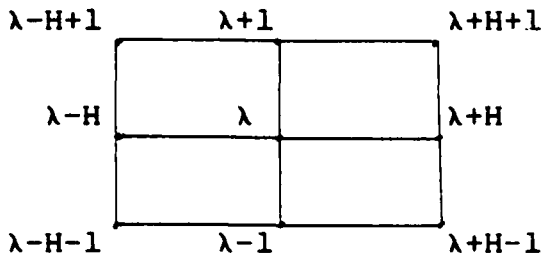


Fig 2 - the neighbors of node λ

$$\begin{array}{lll} \rho_{-4}(\lambda) = \lambda - H - 1 & ; & \rho_4(\lambda) = \lambda + H + 1 \\ \rho_{-3}(\lambda) = \lambda - H & ; & \rho_3(\lambda) = \lambda + H \\ \rho_{-2}(\lambda) = \lambda - H + 1 & ; & \rho_2(\lambda) = \lambda + H - 1 \\ \rho_{-1}(\lambda) = \lambda - 1 & ; & \rho_1(\lambda) = \lambda + 1 \\ \rho_0(\lambda) = \lambda & & \end{array} \qquad (4)$$

Similar neighboring functions may be defined for other discretizations, and then used to determine the stripe structure of the corresponding matrix as illustrated by the following theorem:

Theorem 1: Let the numbering of M_Ω be deduced from that of M_Q, and let A be the stiffness matrix that results from a specific discretization of a PDE on M_Ω (with a specific definition of neighboring nodes). If there exists π functions $\rho_k: M_Q \to M_Q$, $k=1,\ldots,\pi$ such that for any two neighboring nodes λ and μ in M_Q, $\rho_k(\lambda)=\mu$, for some k, $1 \le k \le \pi$, and the functions ρ_k satisfy

$$\rho_k(\lambda) < \rho_k(\lambda+1) \qquad k=1,\ldots,\pi \qquad (5.a)$$
$$\rho_k(\lambda) < \rho_{k+1}(\lambda) \qquad k=1,\ldots,\pi-1 \qquad (5.b)$$

Then, it is possible to construct a stripe structure for A with stripe count π.

Proof: Define, for $k=1,\ldots,\pi$, the following sets:

$$S_k = \{(\ell, \sigma_k(\ell)) \; ; \; 1 \le \ell \le n_\Omega \text{ and } \sigma_k(\ell) \downarrow\} \qquad (6)$$

where

$$\sigma_k(\ell) = \begin{cases} \nu(\rho_k(\nu^{-1}(\ell))) & \text{if } \rho_k(\nu^{-1}(\ell)) \in M_\Omega \\ \uparrow & \text{otherwise} \end{cases}$$

Where $\downarrow$ and $\uparrow$ are used for "defined", and "not defined",

respectively. It is readily seen that if the $(\ell,m)^{th}$ element of A is non zero, then nodes $\nu^{-1}(\ell)$ and $\nu^{-1}(m)$ are neighbors, and there exists a k such that $\nu^{-1}(m)=\rho_k(\nu^{-1}(\ell))$. Thus, $(\ell,m)\epsilon S_k$. In other words, every position of A that has a non-zero element is in some set S_k, $1\leq k\leq\pi$.

In order to prove that each set S_k, $1\leq k\leq\pi$, is a stripe, we consider any two elements $(\ell,\sigma_k(\ell))$ and $(m,\sigma_k(m))$ in S_k. If $\ell=\nu(\lambda)$ and $m=\nu(\mu)$, then by the definition of σ_k, both $\rho_k(\lambda)$ and $\rho_k(\mu)$ are in M_Ω, and hence both $\nu(\rho_k(\lambda))$ and $\nu(\rho_k(\mu))$ are defined. Now if $\ell > m$, then from (3.a) $\lambda > \mu$ and from (5.a) $\rho_k(\lambda) > \rho_k(\mu)$. Thus $\nu(\rho_k(\lambda)) > \nu(\rho_k(\mu))$, that is $\sigma_k(\ell) > \sigma_k(m)$, which proves that σ_k is a strictly increasing function and that S_k is a stripe.

Finally, we need to show that $S_k < S_{k+1}$. For this, we consider the two elements $(\ell,\sigma_k(\ell)) \epsilon S_k$, and $(m,\sigma_{k+1}(m)) \epsilon S_{k+1}$. Following the same steps as above, we may show that if $\ell \geq m$, then $\lambda \geq \mu$ and $\rho_{k+1}(\lambda) \geq \rho_{k+1}(\mu)$. But from (5.b), $\rho_{k+1}(\mu) > \rho_k(\mu)$, which leads to $\sigma_{k+1}(\ell) > \sigma_k(m)$. []

Note that the above theorem does not depend on the specific numbering of M_Ω. For column wise numbering, the functions (4) may be used (assuming $H > 2$) to prove the following:

Corollary 1: If the nodes in a pierced rectangular grid M_Ω are numbered column-wise, then the matrix that results from FE_4 on M_Ω is a striped matrix with stripe count 9 [].

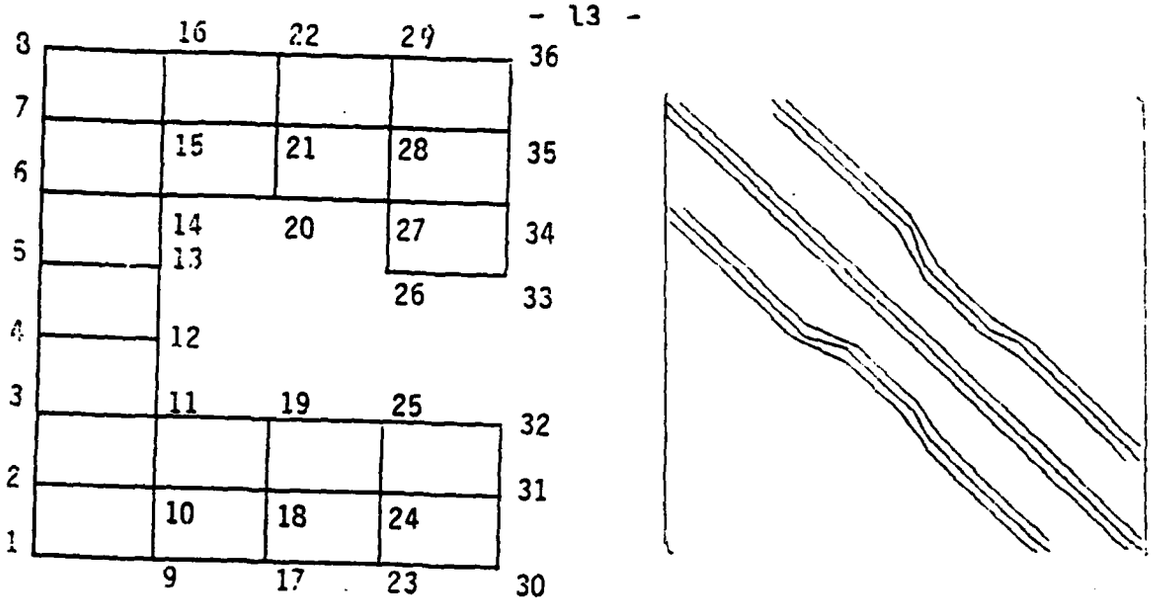
Given that the matrix resulting from FE_4 on M_Q has nine parallel stripes [6], Corollary 1 proves that piercing M_Q and renumbering the nodes do not change the stripe count of the matrix (however, the stripes are no longer parallel).

Results similar to Corrolary 1 may be proved for other discretizations (see table 1 for a summary). Although these results indicate that the network MAT/VEC may be used with the corresponding stiffness matrix, they do not guarantee that the stripes of the matrix are non overlapping, and thus, that the operation of MAT/VEC is not delayed due to internal data conflict. For example, the matrix shown in Fig 3(b), which has overlapping stripes, is obtained from the column wise numbering of the pierced rectangular domain shown in Fig 3(a).

	FD_5	FE_3	FE_4	FE_6	FE_9
regular	5	7	9	19	25
3-color	7(NO)	9(NO)	11(NO)	23	29
5-color	x	x	x	23(NO)	29(NO)

*) NO = Not overlapping if Lemma 2 applies

Table 1 - Stripe count for different numbering schemes



(a) The grid (b) The FE_4 matrix

Fig 3 - column-wise numbering

4. MULTI-COLOR NODE NUMBERING

Many multicolor numbering schemes have been used by different authors to obtain stiffness matrices that have some desirable properties (see e.g. [1,7,8]). In this section, we introduce a multicolor scheme that spreads apart the stripes of A such that they do not overlap. we consider only 3-color numbering, and we assume that H=3h-1, for some integer h. This may be satisfied, always, by increasing the height of M_Q appropriately.

In order to explain the 3-color numbering scheme, we assume that each horizontal line in M_Q is given a color. Namely, lines 1,4,...,3h-2 are white, lines 2,5,...,3h-1 are black, and lines 3,6,...,3h-3 are red. Numbers are, then, assigned to the nodes in M_Q as follows:

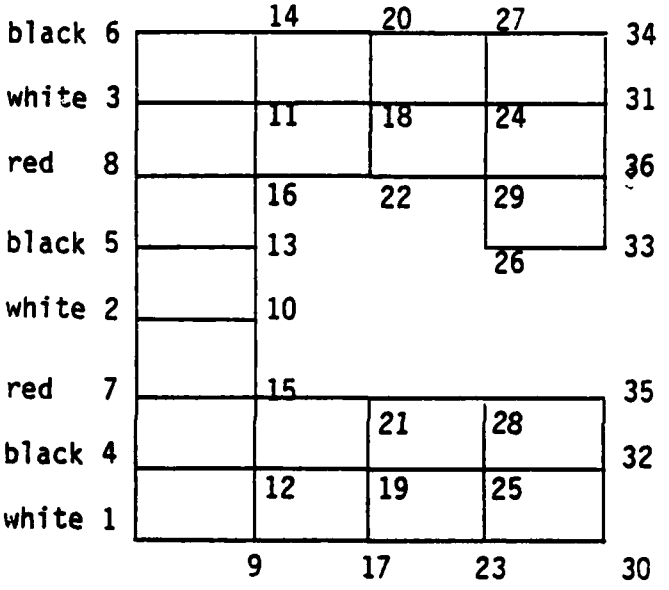
For each column j=1,...,W Do

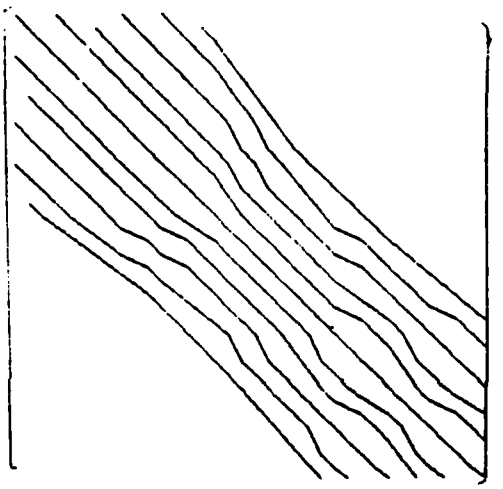
1) number the nodes in the white lines of column j,

2) number the nodes in the black lines of column j,

3) number the nodes in the red lines of column j,

The 3-color numbering of the nodes of M_Ω is then the numbering deduced from the 3-color numbering of M_Q. As an example, we show in Fig 4(a) the 3-color numbering of the same grid of Fig 3(a).



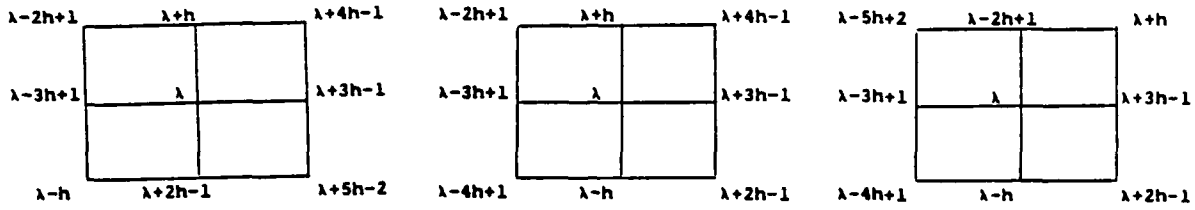


(a) The grid (b) The FE_4 matrix

Fig 4 - 3-colors numbering scheme

As we did for regular numbering, we assume that FE_4 discretization is used and we introduce the appropriate neighboring functions in M_Q. However, in this case, the neighbors of a given node λ depend on the color of λ. In order to be more specific, we show in Fig 5 the numbers that are assigned by the 3-color numbering to the nine neighbors of λ. Clearly, at least eleven functions are needed in order to include all the neighbors. Namely:

$$\begin{array}{lll} \rho_5(\lambda) = \lambda + 5h - 2 & ; & \rho_{-5}(\lambda) = \lambda - 5h + 2 \\ \rho_4(\lambda) = \lambda + 4h - 1 & ; & \rho_{-4}(\lambda) = \lambda - 4h + 1 \\ \rho_3(\lambda) = \lambda + 3h - 1 & ; & \rho_{-3}(\lambda) = \lambda - 3h + 1 \\ \rho_2(\lambda) = \lambda + 2h - 1 & ; & \rho_{-2}(\lambda) = \lambda - 2h + 1 \\ \rho_1(\lambda) = \lambda + h & ; & \rho_{-1}(\lambda) = \lambda - h \\ \rho_0(\lambda) = \lambda & & \end{array} \qquad (7)$$



(a) λ is white (b) λ is black (c) λ is red

Fig 5 - The neighbors of λ.

If h≥2 (H≥5), then the functions (7) satisfy the conditions of Theorem 1, and hence, the resulting matrix may be covered by eleven stripes, which is more than the number of stripes resulting from the regular numbering. However, the stripes in this case are non overlapping provided that M_Ω does not have very narrow regions. This condition on M_Ω is better phrased in the following Lemmas:

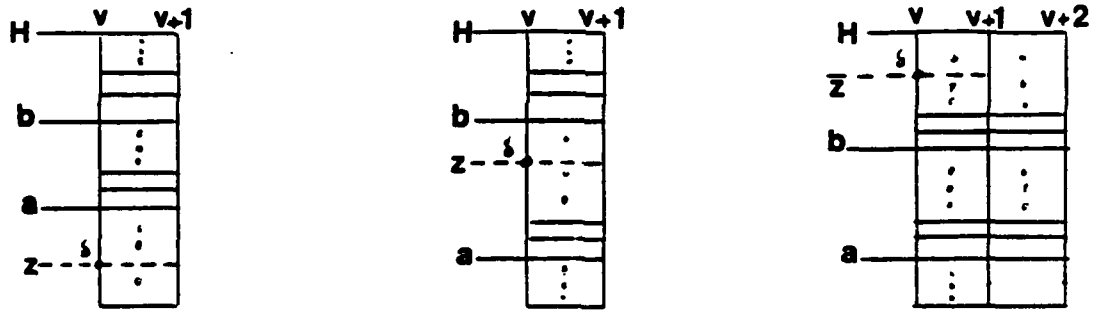
Lemma 1: Assuming 3-color node numbering and 4-nodes quadrilateral elements, if each column in M_Ω contains at least four elements that are either contiguous or divided into two groups of two contiguous elements each, then

$$\text{Next}(\delta) - \delta \leq h - 2 \qquad \text{for any } \delta \in M_\Omega \qquad (8)$$

where the function Next is as given in Definition 2.

Proof: For ease of reference, we indicate the position of a

node that lies on the intersection of the z^{th} horizontal line, $1 \leq z \leq 3h-1$, and the v^{th} vertical line, $1 \leq v \leq W$, of M_Q by the pair (z,v). If $\delta \in M_\Omega$, then Next(δ)=δ and (8) is trivial. Hence, let $\delta \notin M_\Omega$ be at position (z,v) and have the color R. Let also R1 be the color that follows R, that is R1= black, red or white, if R= white, black or red, respectively. From the hypothesis, the v^{th} column of elements in M_Ω contains either four contiguous elements or two pairs of contiguous elements. That is there exists two horizontal lines a and b with $b \geq a+2$ such that all the nodes at positions (c,v) and (c,v+1), for $a \leq c \leq a+2$ and $b \leq c \leq b+2$, are in M_Ω (see Fig 6). Clearly, we may have one of three cases:



Case (1) Case (2) Case (3)

Fig 6 - Column v of M_Ω

Case 1; z < a: In this case, a node $\mu \in M_\Omega$ with color R should exist at a position (c,v), $a \leq c \leq a+2$, and $\mu-\delta-1$ = the number of lines with color R between lines z and a. In other words, $\mu-\delta-1$ is less than the number of lines with color R below line a. Since this number is $\lfloor \frac{a-1}{3} \rfloor$, the largest integer less than $\frac{a-1}{3}$, and given that $b+2 \leq 3h-1$ and $a \leq b-2$, we obtain $\mu-\delta-1 < \lfloor \frac{3h-6}{3} \rfloor = h-2$. By definition, Next($\delta$)$\leq \mu$, and hence Next($\delta$)$-\delta \leq h-2$.

Case 2; $a < z < b$: In this case a node $\mu \epsilon M_\Omega$ with color R should exist at a position (c,v), $b \le c \le b+2$, and $\mu-\delta-1$ is less than the number of lines with color R between lines $a+2$ and b, that is less than $\lfloor \frac{b-(a+2)-1}{3} \rfloor + 1$. Since $b+2 \le 3h-1$ and $a \ge 1$, then $\mu-\delta-1 < \lfloor \frac{3h-7}{3} \rfloor + 1 = h-2$. That is, $Next(\delta)-\delta \le \mu-\delta \le h-2$.

Case 3; $z > b+2$: If $v=W$ and R=red, then, there is no nodes in M_Ω larger than δ. Hence $Next(\delta)=n_Q$, and $Next(\delta)-\delta$ = the number of lines of color R above line $z = \lfloor \frac{3h-1-z}{3} \rfloor$. But $z \ge b+3 \ge a+5 \ge 6$, which gives $Next(\delta)-\delta \le h-3$.

On the other hand, if $v<W$, or $R \neq$ red, then a node $\mu \epsilon M_\Omega$ with color R1 should exist at a position $(c,v+1)$ if R=red or at a position (c,v) if R=white or black, where $a \le c \le a+2$. Here, $\mu-\delta-1$ = [the number of lines of color R above line z] + [the number of lines of color R1 below line a]. But at most $\lfloor \frac{3h-1-z}{3} \rfloor$ lines with color R may be above line z, and at most $\lfloor \frac{a-1}{3} \rfloor$ lines with color R1 may be below line a. Hence $\mu-\delta-1 \le \lfloor \frac{3h-1-z}{3} \rfloor + \lfloor \frac{a-1}{3} \rfloor \le \lfloor \frac{3h-2-z+a}{3} \rfloor$. Given that $z \ge a+5$, then $\mu-\delta-1 \le h-3$, which gives $Next(\delta)-\delta \le h-2$.[]

Lemma 2: Assuming 3-color node numbering and 4-nodes quadrilateral elements. If each column in M_Ω contains at least seven contiguous elements or two separated groups of at least three contiguous elements each, then

$$Next(\delta) - \delta \le h-3 \qquad \text{for any } \delta \epsilon M_Q \tag{9}$$

Sketch of the proof: Let $\delta \notin M_\Omega$ be at position (z,v). Then from the hypothesis, there exists two horizontal lines a and

b, b > a+3 such that all the nodes at positions (c,v) and (c,v+1), for a≤c≤a+3 and b≤c≤b+3 are in M_Ω. The rest of the proof proceeds in a way similar to the proof of Lemma 1. []

Lemma 3: If for any $\delta \in M_\Omega$, Next (δ) - $\delta \leq p$, then,

$$\nu^{-1}(\ell+1) - \nu^{-1}(\ell) \leq p + 1 \qquad \text{for } \ell = 1,\ldots,n_\Omega-1 \quad (10)$$

Proof: let $\ell=\nu(\lambda)$ and $\ell+1=\nu(\bar{\lambda})$. Given that $\bar{\lambda}$ is numbered right after λ, then any node δ with $\lambda<\delta<\bar{\lambda}$ is not in M_Ω, and hence, Next(δ)=$\bar{\lambda}$. But, from the hypothesis, $\bar{\lambda}-\delta\leq p$, and hence, if $\bar{\lambda}>\lambda+1$, then $\bar{\lambda}-\lambda<\bar{\lambda}-\delta\leq p$. That is $\bar{\lambda}-\lambda\leq p+1$. []

The condition (10) may be translated to an upper limit on the number of columns that a stripe in a stiffness matrix may jump in one row. More specifically, if $a_{\ell,\sigma_k(\ell)}$ and $a_{\ell+1,\sigma_k(\ell+1)}$ are in the same stripe, then condition (10) limits the value of $\sigma_k(\ell+1)-\sigma_k(\ell)$. This may be used to prove that the stripes will not overlap if they are adequately separated from each other.

Theorem 2: Let the nodes in a pierced rectangular grid M_Ω be numbered using the 3-color numbering scheme, and let A be the matrix that results from an FE_4 discretization on M_Ω. If M_Ω satisfies the conditions of Lemma 1, then A has eleven non overlapping stripes. Moreover, if M_Ω satisfies the conditions of Lemma 2, then the eleven stripes are strictly non overlapping.

Proof :Consider the eleven functions given by equs (7). It

is straight forward to check that

$$\rho_k(\lambda+1) - \rho_k(\lambda) = 1 \qquad k=-5,\ldots,5 \qquad (11.a)$$
$$\rho_{k+1}(\lambda) - \rho_k(\lambda) \geq h-1 \qquad k=-5,\ldots,4 \qquad (11.b)$$

Clearly, these functions satisfy the conditions in Theorem 1, and hence, A has eleven stripes S_k, $k=-5,\ldots 5$ of the form

$$S_k = \{(\ell,\sigma_k(\ell)) \; ; \; 1\leq\ell\leq n_\Omega \text{ and } \sigma_k(\ell) \downarrow\} \qquad (12)$$

where

$$\sigma_k(\ell) = \begin{cases} \nu(\rho_k(\nu^{-1}(\ell))) & \text{if } \rho_k(\nu^{-1}(\ell)) \in M_\Omega \\ \uparrow & \text{otherwise} \end{cases}$$

In order to prove that these stripes do not overlap, we consider any integers ℓ, m and k such that $(\ell,\sigma_k(\ell))\in S_k$ and $(\ell-m,\sigma_{k+m}(\ell-m))\in S_{k+m}$, and we let $\ell-m = \nu(\bar{\lambda})$ and $\ell = \nu(\lambda)$. From (12), we get

$$\sigma_{k+m}(\ell-m) - \sigma_k(\ell) = \nu(\rho_{k+m}(\bar{\lambda})) - \nu(\rho_k(\lambda)) \qquad (13)$$

But from (11.b) $\rho_{k+m}(\bar{\lambda}) - \rho_k(\bar{\lambda}) \geq m(h-1)$, and from (11.a), $\rho_k(\lambda) - \rho_k(\bar{\lambda}) = (\lambda-\bar{\lambda})$. That is

$$\rho_{k+m}(\bar{\lambda}) - \rho_k(\lambda) \geq m(h-1) - (\lambda-\bar{\lambda})$$

Now, if the conditions of Lemma 1 are satisfied, then from (8) and (10), $\lambda-\bar{\lambda}\leq m(h-1)$ and thus $\rho_{k+m}(\bar{\lambda})-\rho_k(\lambda)\geq 0$. By property (3) of the renumbering function and equation (13), we finally obtain $\sigma_{k+m}(\ell-m) \geq \sigma_k(\ell)$, which is the condition (2) for non overlapping stripes. On the other hand, if the conditions of Lemma 2 are satisfied, then from (9) and (10), $\lambda-\bar{\lambda}<m(h-1)$. This leads to $\sigma_{k+m}(\ell-m) > \sigma_k(\ell)$, which is, the

condition for strictly non overlapping stripes. []

The result of Theorem 2 proves that if an n×n stiffness matrix generated by FE_4 and 3-color numbering is used as input to MAT/VEC, then, execution terminates in n global cycles. In fact, assuming that each y-stream communication line in MAT/VEC may buffer only one data item, the progression of the execution may be described by the following computation fronts:

$$CF_t=\{a_{t-k,\sigma_k(t-k)} \mid -5\leq k\leq 5 \text{ and } \sigma_k(t-k)\downarrow\} \quad t=1,..,n \qquad (14)$$

The 3-color numbering scheme introduced here causes also the stripes of the matrices obtained from FD_5 and FE_3 discretizations to be non overlapping (as indicated in Table 1). However for FE_6 and FE_9 discretizations, this numbering does not spread the stripes enough and overlap may still occur. A 5-color numbering scheme is needed in this case to guarantee non overlapping stripes. The analysis of the 5-color scheme is similar to that discussed in this section.

Although the property of non-overlapping stripes is important for the efficient operation of MAT/VEC, the multi-color numbering scheme has an additional advantage over the regular scheme. Namely, it produces matrices in which the stripes are uniformly spread, thus minimizing the maximum separation between stripes. This is explained in details in the next section.

5. THE MAXIMUM SEPARATION BETWEEN STRIPES

It was shown in [6] that the multiplication of a striped matrix by a vector may be performed on the network MAT/VEC efficiently, only if each communication link directed from a cell k to the previous cell k-1 may buffer at least d_{min} data items, where d_{min} is a measure of the maximum separation between the stripes of the matrix. More specifically, if the stripes of the matrix are full, then d_{min} may be estimated from

$$d_{min} < \max_{k,t}\{\sigma_{k+1}(t) - \sigma_k(t)\}$$

On the other hand, if the stripes of the matrix are not full, then

$$d_{min} = \max_{t,k}\{xP_t(k+1) - xP_t(k)\} \qquad (15)$$

where xP_t, $t=1,\ldots,n$, are the x-stream data profiles corresponding to the execution of MAT/VEC.

In order to observe the effect of the node numbering scheme on the separation between stripes, we consider a rectangular grid M_Q, with H=3h-1 horizontal lines and W vertical lines, and we assume that A is the stiffness matrix that results from FE_4 discretization on M_Q. If regular node numbering is applied, then the functions (4) may be used to construct nine parallel full stripes S_k, $k=-4,\ldots,4$, that satisfy for any t

$$\sigma_{k+1}(t) - \sigma_k(t) = \begin{cases} 1 & \text{if } k=-4,-3,-1,0,2,3 \\ H-2 & \text{if } k=-2,1 \end{cases}$$

This gives $d_{min}<H-2=3(h-1)$. On the other hand, if 3-color numbering is applied, then the functions (7) may be used to produce eleven parallel full stripes S_k, $k=-5,\ldots,5$, that satisfy for any t

$$\sigma_{k+1}(t) - \sigma_k(t) = \begin{cases} h-1 & \text{if } k=-5,-2,1,4 \\ h & \text{if } k=-4,-3,-1,0,2,3 \end{cases}$$

That is $d_{min}<h$. Hence, although the multi-color numbering produces a matrix with a larger band width (5h-1 instead of 3h+1), the stripes are spread within the band almost uniformly, thus decreasing the maximum separation between the stripes from 3(h-1) to h.

The natural question to ask is: does d_{min} remains unchanged if M_Q is pierced and the nodes in the resulting pierced domain M_Ω are renumbered?. More specifically, if we consider the stripe structure discussed in Section 4 and given by (12), can we construct some profile functions that correspond to the fronts (14) such that the maximum in (15) is h?. A positive answer to this question may be provided by considering the following profile functions

$$xP_t(k) = \nu(\text{Next}(\rho_k(\nu^{-1}(t-k)))) \qquad t=1,\ldots,n \qquad (16)$$

where the function Next is defined in Definition 2. Clearly, if $\sigma_k(t-k)\downarrow$, then from (12), $\rho_k(\nu^{-1}(t-k)) \in M_\Omega$, and hence $\text{Next}(\rho_k(\nu^{-1}(t-k))) = \rho_k(\nu^{-1}(t-k))$, which by (12) and (16) gives

$$xP_t(k) = \sigma_k(t-k) \qquad \text{if } \sigma_k(t-k)\downarrow \qquad (17)$$

That is, the knots of the profile (16) coincide with the fronts (14). It is also straight forward to check that

$$xP_t(k) \begin{cases} < xP_t(k+1) & \text{if } \sigma_{k+1}(t-k-1) \downarrow \\ \leq xP_t(k+1) & \text{otherwise} \end{cases}$$

$$xP_t(k) \begin{cases} < xP_{t+1}(k) & \text{if } \sigma_k(t-k) \downarrow \\ \leq xP_{t+1}(k) & \text{otherwise} \end{cases}$$

which are necessary conditions for profile functions [6].

In order to estimate the maximum in (15), we substitute (16) in (15) and get

$$d_{min} = \max_{k,\ell}\{\nu(\text{Next}(\rho_{k+1}(\bar{\lambda}))) - \nu(\text{Next}(\rho_k(\lambda)))\} \qquad (18)$$

where $\lambda=\nu^{-1}(\ell)$ and $\bar{\lambda}=\nu^{-1}(\ell-1)$. Now, let $\phi = \max\{\mu \mid \mu\leq\rho_{k+1}(\bar{\lambda})$ and $\mu\epsilon M_\Omega\}$. That is ϕ is the first node before $\rho_{k+1}(\bar{\lambda})$ that is in M_Ω (take $\phi=0$ if no such μ does exist). Given that $\text{Next}(\rho_{k+1}(\bar{\lambda}))$ is the first node in M_Ω after $\rho_{k+1}(\bar{\lambda})$, we get $\nu(\text{Next}(\rho_{k+1}(\bar{\lambda}))) = \nu(\phi)+1$. Hence

$$\nu(\text{Next}(\rho_{k+1}(\bar{\lambda})))-\nu(\text{Next}(\rho_k(\lambda))) = \nu(\phi)-\nu(\text{Next}(\rho_k(\lambda)))+1 \qquad (19)$$

But $\phi\leq\rho_{k+1}(\bar{\lambda})$ and $\text{Next}(\rho_k(\lambda))\geq\rho_k(\lambda)$. This gives

$$\begin{aligned} \phi - \text{Next}(\rho_k(\lambda)) &\leq \rho_{k+1}(\bar{\lambda}) - \rho_k(\lambda) \\ &\leq \rho_k(\bar{\lambda}) - \rho_k(\lambda) + h \end{aligned} \qquad (20)$$

where we used $\rho_{k+1}(\bar{\lambda})-\rho_k(\bar{\lambda})\leq h$, which may be verified from (7). Also, from (7), $\lambda>\bar{\lambda}$ implies that $\rho_k(\bar{\lambda})< \rho_k(\lambda)$, which together with the property (3.b) of the renumbering function

ν and (20) gives

$$\nu(\phi) - \nu(\text{Next}(\rho_k(\lambda))) < h \qquad (21)$$

From (21) and (19) in (18), we finally get $d_{min} \leq h$. That is, for matrices which are generated from FE_4 discretizations on pierced rectangular grids with hight H and 3-color node numbering, a buffer capacity of $\frac{H}{3}$ is sufficient to ensure that MAT/VEC terminates execution in n global cycles.

6. PERFORMANCE OF MAT/VEC APPLIED TO STIFFNESS MATRICES

As defined in [6], the global cycle of MAT/VEC consists of a communication phase and a processing phase. The time for the processing phase is the time needed to complete a floating point multiply/add, which is constant for a given architecture of the cells of MAT/VEC. On the other hand, the time for the communication phase depends on the stripe structure of the matrix. More specifically, given the stripe structure of the input matrix and a corresponding data profile, the time for the communication phase of the t^{th} global cycle, $1 \le t \le n$, is the time for ξ_t data transmission, where

$$\xi_t = \max_k \{xP_t(k) - xP_{t-1}(k)\} \tag{22}$$

Assuming that the time needed to complete a multiply/add operation is τ_m, and the time needed to transmit a single data item between two cells is τ_c, then the total execution time of MAT/VEC is

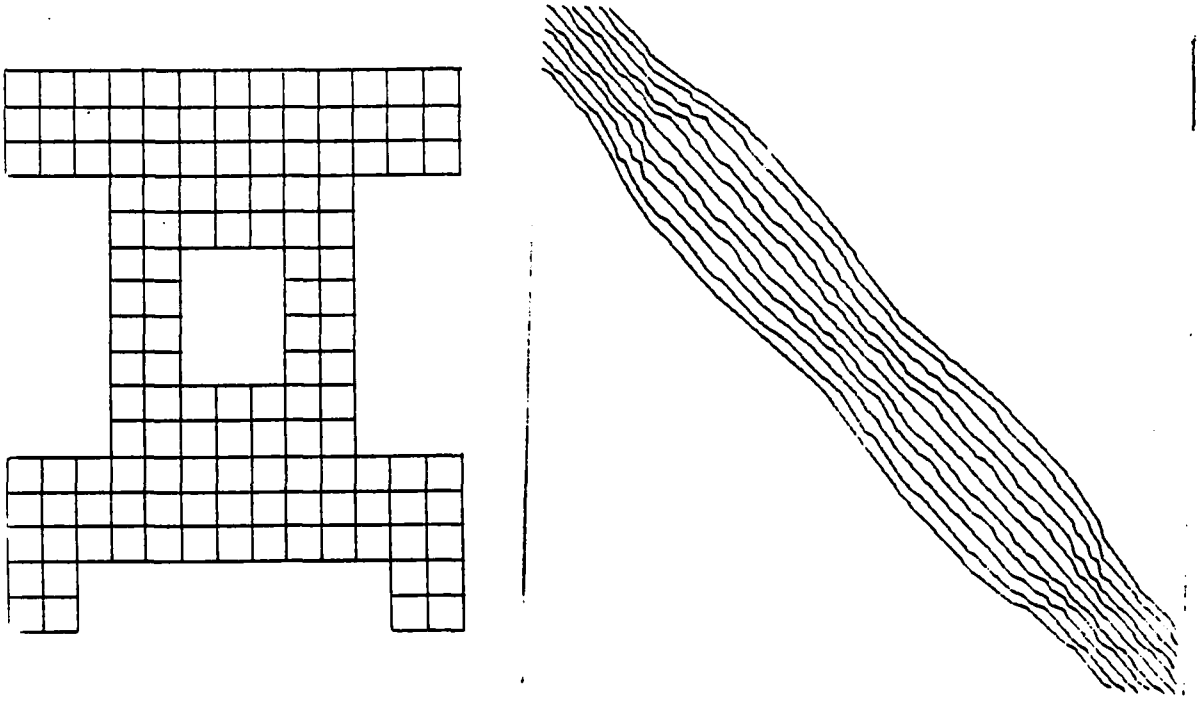
$$T = \tau_m + \tau_c \sum_{t=1}^{n} \xi_t \tag{23}$$

For stiffness matrices resulting from 3-color numbering and FE_4 discretizations, it is easy to show that the profile functions (16) lead to

$$\xi_t \le h \qquad t=1,\ldots,n$$

However, the actual value of ξ_t is usually much smaller than h, as shown by the following example.

EXAMPLE 1:



(a) The grid (b) The corresponding matrix

Fig 7 - Example 1

Consider the pierced rectangular grid shown in Fig 7(a). It contains 130 4-nodes rectangular elements and 174 nodes. The stiffness matrix corresponding to the 3-color numbering scheme (Fig 7(b)) has a band width b=49 and, in accordance with Theorem 2, has 11 strictly non-overlapping stripes. The construction of the data profiles (16) and the application of (22) gives $\sum_{t=1}^{174} \xi_t = 241$. That is MAT/VEC completes the multiplication of the matrix by a vector in time $T = 174\tau_m + 241\tau_c$. Note that if the multiplication is performed on a systolic network [5], then 49 cells are needed and the computation is completed in time $T_s = 358(\tau_m + \tau_c)$. The saving in both the number of cells and the execution time is

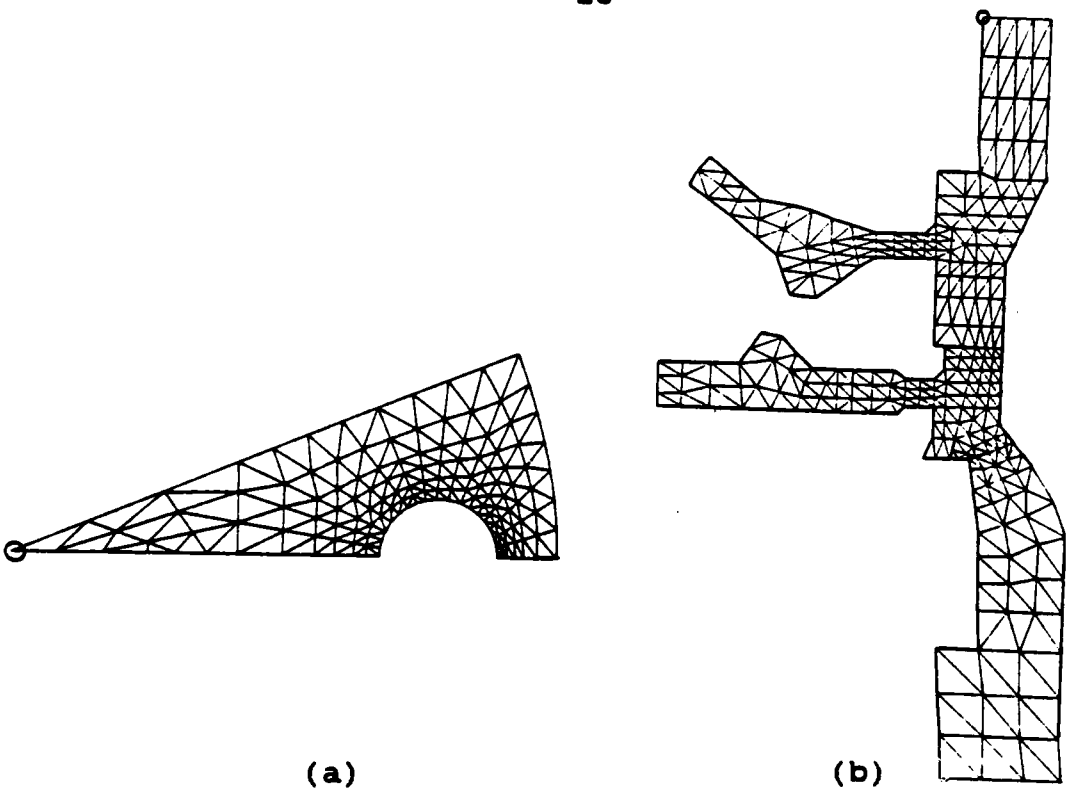
obvious. Also, the number of cells π in MAT/VEC is independent of the size of the grid, while the number of cells used in the systolic approach [5], namely b, depends on the size of the grid (usually, $b = O(\sqrt{n})$).

In order to observe the effect of the numbering scheme, we also consider the matrix corresponding to the column-wise numbering of the grid of Fig 7(a). This matrix has a bandwidth $b_n = 35$ and may be covered by nine stripes. However, the stripes are not "strictly non overlapping", and the construction of the computation fronts (see [6]) shows that 326 global cycles are needed for the completion of the execution of MAT/VEC. Hence, the size of MAT/VEC is smaller for regular numbering than in the 3-color numbering (9 cells instead of 11 cells), but execution is slower (326 global cycles instead of 174 cycles).

Clearly, general results, of the type proved in the previous sections, may only be obtained for grids that are, in some sense, regular. However, given any sparse matrix, and in particular a stiffness matrix, a stripe structure may be constructed for the matrix and the number of computation fronts needed for the execution of MAT/VEC may be estimated.

EXAMPLE 2:

Highly irregular grids may be obtained if triangular elements are used. Consider, for example, the two grids shown in Fig 8 that are extracted from [9].The Cuthill-Mckee



(a) (b)

Fig 8 - Irregular grids

numbering scheme [2] is used for both grids starting from the encircled nodes. The stiffness matrix corresponding to the grid of Fig 8(a) is of order 145 and has a band-width 25. The minimum number of stripes that may cover the matrix is 9 (overlapping) and the number of computation fronts is found to be 283 fronts. For the grid of Fig 8(b), the order of the matrix is 289 and the bandwidth is 49. The number of stripes is found to be 13 and the corresponding number of computation fronts is 533. By comparision with systolic multiplication, in which all trivial operations are performed, it is clear that the organization of the non-zero elements into a stripe structure, which is independent of the size of the problem, reduces the hardware needed for the completion of the multiplication, without slowing down execution.

Finally, we note that the grids in Fig 8 are

constructed without any consideration for the regularity of the stripe structure. More specifically, the same domains may be easily covered by grids that have the same element-density distribution of the given grids, but that are isomorphic to some pierced rectangular grids. The matrices generated from these grids should obey the results of Section 3 and 4.

7. CONCLUSION

It is shown that the number of stripes π in the stripe structure of a stiffness matrix is independent of the size of the problem, and is much smaller than the band-width of the matrix. For pierced rectangular domains, the stripe count π may be estimated analytically and the stripe structure of the matrix may be constructed from the finite element grid.

The multicolor node numbering presented in this paper has two favorable effects on the resulting matrix: First, it produces non-overlapping stripes, which prevents any data conflict during the execution of MAT/VEC, and second, it distributes the stripes uniformly, which reduces the maximum separation between stripes and thus minimizes the number of buffers needed in MAT/VEC.

In brief, the construction of stripe structures for stiffness matrices allows for the efficient utilization of VLSI networks. Moreover, the number of cells in such networks is determined by the stripe count π, which is independent of the size of the problem.

REFERENCES

[1] L. Adams, "Iterative Algorithms for Large Sparse Linear Systems on Parallel Computers," Ph.D. Thesis, Univ. of Virginia (October 1982).

[2] E. Cuthill and J. Mckee, "Reducing the Bandwidth of Sparse Symmetric Matrices," Proccedings of the ACM National Conf, New-York (1969), pp. 157-172.

[3] S. Eisenstat, M. Gursky, M. Schultz and A. Sherman, "Yale Sparse Matrix Package," Tech. Report 112,114, Dept. of Computer Science, Yale University (1977).

[4] A. George and J. Liu, "Computer Solutions of Large Sparse Positive Definite Systems," Prentice-Hall series in Computational Mathematics (1981).

[5] H. T. Kung and C. E. Leiserson, "Systolic Arrays for VLSI," in Introduction to VLSI Systems (1980). Ed. by C. Mead and L. Conway, Addison-Wesley, Reading, Mass.

[6] R. Melhem, "Parallel Solution of Linear Systems with Striped Sparse Matrices; Part 1: VLSI Networks for Striped Matrices," Tech. Report. ICMA-86-91 (January 1986).

[7] D. O'leary, "Ordering Schemes for Parallel Processing of Certain Mesh Problems," SIAM J. on Sci. and Stat. Computing, Vol 5-3 (Sept 1984), pp. 620-632.

[8] Y. Saad and M. Schultz, "Parallel Implemetations of Preconditioned Conjugate Gradient Methods," Tech. Report YALEU/DCS/RR425, Dept. of Computer Science, Yale University (Oct. 1985).

[9] O. C. Zienkiewicz, "The Finite Element Method," McGraw-Hill, (1979).

END

DTIC

5-86